Segmental isotopic labeling by asparaginyl endopeptidase-mediated protein ligation

Mikula, Kornelia M.

2018-08

Mikula , K M , Krumwiede , L , Plueckthun , A & Iwai , H 2018 , ' Segmental isotopic labeling by asparaginyl endopeptidase-mediated protein ligation ' , Journal of Biomolecular NMR , vol. 71 , no. 4 , pp. 225-235 . https://doi.org/10.1007/s10858-018-0175-4

http://hdl.handle.net/10138/324063
https://doi.org/10.1007/s10858-018-0175-4

acceptedVersion

Segmental isotopic labeling by asparaginyl endopeptidase-mediated protein ligation

Kornelia M. Mikula[1], Luisa Krumwiede[1], Andreas Plückthun[2], & Hideo Iwaï [1,*]

[1]*Research Program in Structural Biology and Biophysics, Institute of Biotechnology, University of Helsinki. P.O. Box 65, Helsinki, FIN-00014, Finland*

[2]*Department of Biochemistry, University of Zürich, Winterthurerstrasse 190, 8057, Zürich, Switzerland.*

*To whom correspondence should be addressed

Phone: +385 9 191 59752

Email: hideo.iwai@helsinki.fi

Keywords: asparaginyl endopeptidase, designed armadillo repeat protein, segmental isotopic labeling, protein *trans*-ligation, protein NMR

Abbreviations: PTS, protein-*trans* splicing; CTS, conditional protein *trans*-splicing; EPL, expressed protein ligation; IPL, intein-mediated protein ligation; NCL, native chemical ligation; SML, sortase-mediated ligation; AEP, asparaginyl endopeptidase; AML, AEP-mediated ligation; SrtA, *Staphylococcus aureus* sortase A; HSQC, heteronuclear single-quantum correlation spectroscopy; *Oa*AEP1, *Oldenlandia affinis* asparaginyl endopeptidase 1_b; HINT domain, hedgehog/intein domain; dArmRP, designed armadillo repeat protein; GFP, green fluorescent protein; MBP, maltose-binding protein; Ulp1, Ubiquitin-like-specific protease; IMAC, immobilized metal ion affinity chromatography; IPTG, isopropyl β-D-1-thiogalactopyranoside.

ABSTRACT

Segmental isotopic labeling can facilitate NMR studies of large proteins, multi-domain proteins, and proteins with repetitive sequences by alleviating NMR signal overlaps. Segmental isotopic labeling also allows us to investigate an individual domain in the context of a full-length protein by NMR. Several established methods are available for segmental isotopic labeling such as intein-mediated ligation, but each has specific requirements and limitations. Here, we report an enzymatic approach using bacterially produced asparagine endopeptidase (AEP) from *Oldenlandia affinis* for segmental isotopic labeling of a protein with repetitive sequences, a designed armadillo repeat protein, by overcoming some of the shortcomings of enzymatic ligation for segmental isotopic labeling.

Introduction

Segmental isotopic labeling is a powerful labeling technique in protein NMR spectroscopy to facilitate NMR studies of larger proteins by not only alleviating signal overlaps in larger proteins but also preserving features of uniformly isotope-labeled samples for triple resonance experiments (Skrisovska et al. 2010; Volkmann & Iwaï 2010). Several approaches have been successfully applied to produce segmentally isotope-labeled proteins. Segmental isotopic labeling methods can be divided into three categories: (1) Chemoselective ligation, (2) Protein *trans*-splicing (PTS), and (3) Protease-mediated ligation. Each method has some specific requirements and limitations. Chemoselective ligation such as the Expressed Protein Ligation (EPL)/Intein-mediated Protein Ligation (IPL) requires an α-thioester group at the C-terminus and an N-terminal cysteine residue (Fig. 1A) (Dawson et al. 1994; Evans et al. 1998; Muir et al. 1998; Xu et al. 1999). PTS requires fusions with a split intein or HINT (hedgehog/intein) domains, which could reduce the solubility of split fragments and might require refolding steps (Fig. 1B) (Yamazaki et al. 1998; Otomo et al. 1999a, b; Muona et al. 2010; Aranko et al. 2013; Aranko et al. 2014). However, PTS can also be used for segmental isotopic labeling of a central fragment by multiple-fragment ligation steps using two orthogonal split inteins (Otomo et al. 1999b; Busche et al. 2009). Enzymatic ligation for segmental labeling typically requires specific recognition sequences in the substrates and an enzyme (Mao et al. 2004; Kobashigawa et al. 2009; Freiburger et al. 2015).

Despite its potential usefulness, the applications of segmental isotopic labeling have been relatively limited. This is because segmental isotopic labeling is often seen as technically demanding and labor-intensive, as well as requiring undesired amino-acid changes for ligation. For example, multiple-step protein purifications of at least two fragments are required together with optimizations for the ligation conditions to generate the ligated protein (Otomo et al. 1999a; Skrisovska et al. 2008; Minato et al. 2012).

There have been several improvements reported for each of these methods. PTS can also be performed *in vivo* without multiple protein purification steps (Züger & Iwai 2005; Muona et al. 2010). Ligation by EPL can also be improved by different thiol reagents or on-column ligation (Johnson & Kent 2006; Skrisovska et al. 2008;

Michel et al. 2013; Gallagher et al. 2017). Conditional splicing using a salt-inducible intein has been recently exploited to alleviate the solubility issue of split inteins (Fig. 1C) (Ciragan et al. 2016). Enzymatic ligation using Sortase A (SrtA) from *Staphylococcus aureus* is an attractive approach because it requires only short tags instead of larger split intein fragments (Mao et al. 2004). Sortase-mediated ligation (SML) has increasingly become popular in protein modification (Fig. 1D) (Mao et al. 2004; Antos et al. 2016). However, Sortase-mediated ligation has strict sequence requirements and thus does not permit arbitrary ligation sites (see below).

For segmental isotopic labeling of proteins, the major challenge is to achieve both high ligation efficiency and very high purity, which is required for NMR studies (Muona et al. 2010; Minato et al. 2012). One of the drawbacks of any enzymatic ligation, including SML, could be the reverse reaction catalyzed by the enzyme because the ligated product still contains the recognition sequence. The reverse reaction could be suppressed by constantly removing one of the products (Freiburger et al. 2015). Moreover, SrtA is a relatively inefficient enzyme with the low turnover rate. Therefore, a large amount (stoichiometric amount) of the enzyme and/or longer reaction time is often used to produce a large quantity of ligated proteins required for NMR investigations (Kobayashikawa et al. 2009; Freiburger et al. 2015). Variants of SrtA have been developed to improve the transpeptidase kinetics of SrtA (Dorr et al. 2014). However, another shortcoming of SrtA is an insertion of "Leu-Pro-Xxx-Thr-Gly-Gly" (LPXTGG) sequence at the ligation site, which is necessary for SrtA to ligate. The recognition sequence introduces a non-natural sequence and could thus affect biological functions of ligated proteins. Therefore, it might not be the best choice for NMR studies of biologically active proteins.

In contrast, the PTS approach using an intein bearing Ser at the +1 position has been successfully demonstrated to be valuable for producing a segmentally isotope-labeled protein with the native sequence, which can be accommodated owing to the sequence diversity of naturally occurring inteins and HINT domains (Ciragan et al. 2016).

Recently, we demonstrated segmental isotopic labeling of a single-domain globular protein without any refolding step using asparaginyl endopeptidase (AEP) from *Oldenlandia affinis* (*Oa*AEP1) (Mikula et al. 2017). AEPs from plants catalyze an efficient transpeptidase reaction in backbone cyclization of cyclotides (Fig. 2)

(Nguyen et al. 2014; Harris et al. 2015). Importantly, *Oa*AEP1 was produced from an *E. coli* bacterial expression system similar to SrtA, making it an attractive alternative enzyme for segmental isotopic labeling (Harris et al. 2015; Mikula et al. 2017). The natural recognition sequences of *Oa*AEP1 are minimal, "Gly-Leu"(GL) and "Thr-Arg-Asn/Gly-Leu" (TRN/GL), in the natural substrate of Kalata B1 precursor (a cyclotide) for the N- and C-termini respectively, in which "/" indicates the cleavage and ligation site (Fig. 1E) (Harris et al. 2015). Additionally, *Oa*AEP1 seems to be tolerant of variations in the sequence, even though the full range of tolerable mutations is still unknown. This suggests that the practical sequence requirement can even be shorter in length with the theoretical requirement of only one residue of "Asn" (Harris et al. 2015; Mikula et al. 2017), but the scope of possible sequence changes remains to be investigated. We previously reported that proteins with the sequences of "TRN/AL" or "TRN/CL" in the C-terminal propeptide sequence were much more efficient substrates than the original "TRN/GL" for recombinant *Oa*AEP1 at least for the backbone cyclization of a model protein of green fluorescent protein (GFP) (Mikula et al. 2017).

Here, we report the use of asparagine endopeptidase (AEP) for production of segmental isotope-labeled proteins by overcoming the shortcoming of AEP-mediated ligation for complementing existing segmental isotopic labeling techniques and discuss the possible applications and limitations.

Materials and methods

GFP substrates for trans*-ligation*

The N-terminal substrate GSH_6-(GFP)-TRNCL was produced from the vector pEMRSF2, derived from pJTRSF50 using two oligonucleotides, I940: 5'- ACA TAT GGG CAG CCA TCA TCA CCA TCA CCA C and J052: 5'- AGG AAG CTT ACA GAC AAT TTC GGG TAC TAC CGC G (Muona et al. 2010; Mikula et al. 2017). The C-terminal substrate of $GLPH_6$-(GFP(N159S))-GLT was produced as a SMT3 fusion protein (the yeast SUMO protein) from the plasmid pBHRSF277 bearing an N159S mutation in GFP introduced by using the two oligonucleotides, J166: 5'-GAA GTT AGC TTT GAT CCC ACT CTT TTG TTT GTC TGC and J167: 5'- GCA GAC

AAA CAA AAG AGT GGG ATC AAA GCT AAC TTC and pBHRSF208 as the template. The protein was expressed as a SUMO fusion and purified as previously described (Guerrero et al. 2015; Mikula et al. 2017).

Trans-*ligation using model globular proteins*

Purified substrate proteins were mixed at an equimolar ratio at a final concentration of 5 μM each in the reaction buffer of 50 mM HEPES, pH 7.0, 50 mM NaCl, 1 mM EDTA, and 0.5 mM TCEP. Pre-activated recombinant *Oa*AEP1, of which self-proteolysis was induced for the activation by lowering pH of the buffer to 4.0 for 5 hours, was added at a final concentration of 0.2 μM to GFP variants. Reactions were incubated at room temperature in 1.5 mL Eppendorf tubes. The samples for SDS-PAGE analysis were taken every 10 minutes for 100 minutes for GFP variants. The time course of the *trans*-ligation of two GFP variants was analyzed by SDS-PAGE. The samples taken at 10-minute intervals for 100 minutes were diluted 1:1 with 2-times SDS loading buffer, and loaded on 12% SDS polyacrylamide gels. The SDS-gels were stained with Coomassie Blue R (GE Healthcare) and quantified using NIH ImageJ software with the assumption that the dye binds equally to the proteins (Schneider et al. 2012).

Segmental isotopic labeling of designed armadillo repeat protein (dArmRP)

The full-length dArmRP (YM_3A) consists of the N-terminal capping repeat (Y), three consensus repeats (M) and a C-terminal artificially designed capping repeat (A). The gene of YM_3A was split into two fragments of YM and MMA with the N- and C-terminal additions as below. The N-terminal fragment was designed to have the C-terminal sequence of "NCL" for ligation by *Oa*AEP1. The N-terminal fragment was cloned into pHYRSF53 as H_6-SUMO fusion at the N-terminus, using the two oligonucleotides, HK542: 5'-AAG GAT CCG AAC TGC CGC AGA TGA C and J153: 5'- GGA AGT GTT GTT TCA GGG TCC TGA TGG CAA CGC GCT GAC CC, resulting in pEMRSF9 bearing H_6-(SMT3)-(YM) with "NCL" at the C-terminus. The gene of the C-terminal fragment of MMA with the N-terminal sequence of "GL" for ligation was amplified by PCR using the two oligonucleotides, J157: 5'- ACA TAT GGG TCT TAA CGA ACA AAT CCA AG and J223: 5'- TTA TGA ATT CGT

GGG ACT GCA GCT TCT C. The PCR product was cloned into pMHBAD14, resulting in pEMBAD26 bearing GL-(M_2A)-H_6 (Parmeggiani et al. 2008; Oeemig et al. 2009; Alfarano et al. 2012).

For segmental isotopic labeling, the N-terminal fragment (YM) was expressed in 20% ^{13}C, 100% ^{15}N-labeled M9-medium containing $^{15}NH_4Cl$ (1 g/L) and a mixture of $^{13}C_6$ D-glucose (0.2 g/L) and unlabeled D-glucose (0.8 g/L) as nitrogen and carbon sources. T7 Express Competent *E. coli* cells (New England Biolabs) were transformed with the plasmid (pEMRSF9) for protein expression. The cells were grown at 37 °C in 2 liters of M9-medium supplemented with 25 μg/ml kanamycin and induced with a final concentration of 1 mM of isopropyl β-D-1-thiogalactopyranoside (IPTG) when the OD_{600} reached 0.6. The cells were further incubated for 3 hours at 200 rpm before harvesting. The protein was purified by immobilized metal ion affinity chromatography (IMAC) using a HisTrap 5 ml column (GE Healthcare) as described previously and dialyzed against 2 liters of 20 mM Tris buffer, pH 8.0 and 150 mM NaCl overnight (Guerrero et al. 2015). The C-terminal fragment of GL-(MMA)-H_6 was expressed in the *E. coli* strain ER2566 at 37 °C using pEMBAD26 in 2 liters of LB medium supplemented with 100 μg/ml ampicillin and induced for 3 hours with a final concentration of 0.02% arabinose. The protein was purified by IMAC using a 5 ml HisTrap column and dialyzed against 2 liters of 20 mM Tris buffer, pH 8.0 and 150 mM NaCl, overnight, at 8 °C.

The labeled N-terminal fragment (YM)-NCL (133 μM) and the unlabeled C-terminal fragment GL-(MMA)-H_6 (67 μM) was mixed at a 2:1 molar ratio in a volume of 1.5 mL. The ligation reaction was initiated by addition of 420 μl of the pre-activated recombinant *Oa*AEP1, of which proteolytic activity was activated by lowering pH to 4.0 for 5 hours, at a final concentration of 2.7 μM. The preparation of activated recombinant *Oa*AEP1 using the plasmid pBHRSF184 (Addgene ID #89482) was previously reported (Mikula et al. 2017). The reaction mixture was immediately transferred into a dialysis tube (3.5 kDa MWCO) and dialyzed at room temperature against 0.5 liters of 50 mM HEPES buffer, pH 7.0, 50 mM NaCl, 1 mM EDTA, and 0.5 mM TCEP overnight. The reaction mixture was further purified by anion exchange chromatography using a MonoQ™ 5/50 GL column (GE Healthcare Life Sciences). The fractions containing the ligated product were collected and dialyzed

overnight at 8 °C against 1 liter of 20 mM sodium phosphate buffer, pH 6.0 and concentrated to 0.58 mM using an ultracentrifugation device.

For the comparison, a full-length protein was also produced in the labeled M9 medium using the plasmid (pADHRSF57) following the same purification protocol and concentrated to 0.45 mM in 20 mM sodium phosphate buffer, pH 6.0.

Amino acid tolerance of Oa*AEP1at the N-terminal P1" and P2" sites*

To test the promiscuity of *Oa*AEP1 at the N-terminus (P1" site in Fig. 6a), we prepared GFPs with three different amino acids (Gly, Ala, and Ser) at the N-terminus. GLP-(GFP)-RNALPH$_6$, ALP-(GFP)-RNALPH$_6$, SLP-(GFP)-RNALPH$_6$ were produced as the N-terminal SUMO-fusion proteins from the plasmids of pJTRSF82, pKERSF10, and pKERSF11, respectively. The removal of the N-terminal SUMO domain in these fusion proteins by Ulp1 protease created the N-terminal Gly, Ala, or Ser. pJTRSF82 was derived from pJTRSF55 using the following oligonucleotides, #78GFP_N: 5'- TGG GAT CCA AAG GAG AAG AAC NTT NC and J164: 5'-GAT GAT GAT GAT GAT GAC CCA GAG CAT TTC GGG TAC T, and HK122: 5'-CTA AAG CTT AAT GAT GAT GAT GAT GAT G to add the C-terminal His$_6$ tag. pKERSF10 and pKERSF11 were constructed using the following oligonucleotides, J136: 5'-TAG GTA CCC GGC AGT GCT CCA CCA ATC TGT TCT C and J138: 5'-TAG GTA CCC GGC AGT GAT CCA CCA ATC TGT TCT C, respectively and cloned between the *Spe*I and *Kpn*I sites of pJTRSF82. For testing the mutations at the P2" position, GVP-(GFP)-RNALPH$_6$, GIP-(GFP)-RNALPH$_6$, GFP-(GFP)-RNALPH$_6$, GMP-(GFP)-RNALPH$_6$, were produced as the N-terminal SUMO-fusion proteins from the plasmids of pJTRSF125, pJTRSF127, pJTRSF128, and pJTRSF129, respectively. JTRSF125, pJTRSF127, pJTRSF128, and pJTRSF129 were created by PCR using the following oligonucleotides containing the mutations, J290: 5'- TAG GTA CCC GGA ACT CCT CCA CCA ATC TGT, J292: 5'-TAG GTA CCC GGG ATT CCT CCA CCA ATC TGT, J293: 5'-TAG GTA CCC GGC ATT CCT CCA CCA ATC TGT and JT010: 5'-TAG GTA CCC GGG AAT CCT CCA CCA ATC TGT, respectively.

The protein expression was induced at 22 °C and proteins were purified as described previously (Mikula et al. 2017). Cyclization reactions of GFP variants were carried out, of which three varied the N-terminal residue ('GLP', 'ALP', and 'SLP')

and four varied the second residue (the P2" position in Fig. 6A; 'GVP', 'GIP', 'GMP, and 'GFP') together with Ala at the P1' position (Fig. 6). These were performed using 5 μM of the substrate and 0.2 μM of activated recombinant *Oa*AEP1 and analyzed as previously reported (Mikula et al. 2017).

NMR measurements

[^{1}H, ^{15}N]-HSQC spectra were recorded on a Bruker Avance III HD spectrometer equipped with a cryogenically cooled probe at a ^{1}H frequency of 850 MHz. The segmentally labeled sample was concentrated to 0.58 mM, 625 μl. The uniformly labeled dArmRP was concentrated to 0.45 mM, 3.8 ml. 450 μl solution containing 5% D_2O of each of the samples was transferred into NMR tubes for the measurements.

RESULTS

***Trans*-ligation of two globular domains without any affinity for each other by *Oa*AEP1**

Efficient enzymatic ligation of two globular domains with a peptide bond would make it very simple to introduce segmental isotopic labeling for multi-domain proteins. Robust enzymatic activities of backbone cyclizing AEPs from plants are very attractive for the protein ligation, compared to widely used SrtA. AEPs would require a smaller amount of enzyme and the sequence modification can be a half or fewer in length of what is used with SrtA. Despite the reaction is much faster than SrtA with smaller enzyme-to-substrate ratios, the ligation efficiency of two globular proteins (GFPs) without any affinity for each other to create a tandem fusion was not impressive (Fig. 3). The low efficiency is presumably due to the reverse reaction (Fig. 2A). Therefore, it might be unlikely to achieve a high ligation yield (>90%) (Fig. 3C). This could restrict the application of AML for *trans*-ligation of multiple domains when the ligation efficiency is critical, e.g., segmental isotopic labeling. The shorter recognition sequence of *Oa*AEP1 might also lead to non-specific cleavages because possible non-specific degradation was previously observed when a small domain of the B1 domain of IgG binding protein G (GB1) was ligated to GFP (Mikula et al. 2017). Thus, smaller less stable proteins and flexible peptide linkers might not be

suitable as the substrate of *Oa*AEP1. This promiscuity of *Oa*AEP1 might thus limit the application of AML and needs to be further investigated.

Ligation of a designed armadillo repeat protein

Similar to SML, a ligated product of interest by AML remains the substrate for AEP, resulting in the reverse reaction (Fig. 2A). Therefore, the ligation yield might not be very high unless the ligated product is energetically or entropically favorable (Fig. 3). For example, the ligation yield has been improved by continuously removing one of the products during the reaction (Freiburger et al. 2015). As *Oa*AEP1 is very efficient having >200 times faster ligation kinetics than SrtA, such an approach to remove a product might not be very practical. To overcome this problem, the proximity effect can be exploited for improving the ligation efficiency (Fig. 2C). Backbone cyclization of the natural substrate (Kalata B1) of *Oa*AEP1 is very efficient as efficiency close to 100 % was easily achieved (Harris et al. 2015). This is presumably due to the proximity effect by N- and C-termini being closely neighbored in the structure (Fig. 2B).

Previously, we produced a nicked maltose-binding protein (MBP) with AEP-tag by using a dual co-expression system of two fragments that can reassemble, in which the nicked site has N- and C-termini in spatial proximity (Mikula et al. 2017). It was possible to ligate >90% of the nicked MBP by the recombinant *Oa*AEP1 with two residue mutations in the loop by exploiting the natural sequence of "NG" in MBP, supporting the notion that the proximity effect can improve the ligation efficiency, similar to the effect seen in backbone cyclization (Mikula et al. 2017). For avoiding subsequent *in vitro* refolding steps of MBP, time-delayed dual co-expression was used by co-expressing the two split fragments, first in one medium then in the other (differently labeled), to produce the nicked MBP with only one of the fragments labeled (Züger & Iwai 2005; Mikula et al. 2017).

However, not all proteins can be produced as nicked proteins by co-expression of two split fragments. This requirement might constrain the wider applications of this approach. Therefore, we were interested in other protein fragments with inherent affinities *in vitro*. Armadillo repeats are found in many proteins with a repetitive amino acid sequence of about 40 residues in length. Designed armadillo repeat proteins (dArmRP) have been developed as a promising modular scaffold protein for

the engineering of binding molecules that recognize extended polypeptide chains (Parmeggiani et al. 2008; Reichen et al. 2014). Interestingly, dArmRP fragments that have been split between the consensus repeats can spontaneously assemble into one globular protein by non-covalent association (Watson et al. 2014). Thus, this re-associated complex would be an ideal substrate for AEP. Moreover, segmental isotopic labeling is particularly useful for proteins with repeating sequences because NMR signals tend to overlap due to the similar chemical environments within the repeating sequences (Busche et al. 2009). As a model system, we chose a consensus dArmRP (YM_3A) protein, consisting of three identical internal repeats (M), flanked by N- and C-terminal capping repeats (Y and A, respectively). YM_3A was split into two fragments within a loop after the first internal repeat (M_1), namely YM, and M_2A fragments, respectively (Figs. 4A and 5A). We added four residues at the C-terminus of the YM fragment as an AEP-tag, i.e., "KNCL", which is the same sequence we previously used (Fig. 5B) (Mikula et al. 2017). For the C-terminal fragment of M_2A, we created a "GL" sequence at the N-terminus. The ligated product would have three-residue insertion because we utilized a Gly residue in the native sequence (Fig. 5B). We ligated these two fragments *in vitro* by *Oa*AEP1 at a 1:25 enzyme-to-substrate ratio. The ligation was not as fast as the two-GFP ligation, but the ligated product yield exceeded 50% (Fig. 4B). Increasing the amount of YM fragment by two-fold improved the ligated product yield, presumably because the equilibrium was shifted towards the complex. The ligated product yield was >90% after overnight incubation (Fig. 4B).

Segmental isotopic labeling by *Oa*AEP1

Next, we proceeded to produce a segmentally isotope-labeled sample by preparing the N-terminal fragment of YM in the ^{15}N-labeled medium. The solution mixture containing the two purified labeled YM and unlabeled M_2A fragments was simply incubated with the activated recombinant *Oa*AEP1 at an enzyme-substrate-ratio of 1:25 and the product was further purified by anion exchange chromatography to remove unreacted substrates and the enzyme (Fig. 4C, Supplemental Fig. 1). Fig. 5C shows the HSQC spectrum of the segmentally ^{15}N-labeled YM_3A with a three-residue insertion in the YM region (red in Fig. 5). The spectrum with the reduced number of dispersed peaks indicates the well-folded fragment as previously reported, suggesting

the successful segmental isotopic labeling (Watson et al. 2014). We compared the HSQC spectra between the original YM_3A without any ligation and the ligated YM_3A (Fig. 5C and Supplemental Fig. 2) and observed small shifts of several peaks and additional peaks, presumably originating from the newly inserted residues. Even though the amino acid changes are smaller than ones required for SrtA, the insertion in the loop has some influences on the NMR spectrum, and possibly on the three-dimensional structure as well, inducing some chemical shift changes (Supplemental Fig. 2). This observation implies that a small insertion might still influence the functionality of the ligated protein, and it might be necessary to be further minimized for producing biologically active proteins with native primary structures.

Expanding possible ligation sites by *Oa*AEP1

We further asked if we could use *Oa*AEP1 for other sequences because another AEP, butelase 1, was found to be promiscuous at the N-terminal sequence (Nguyen et al. 2014). We mutated the N-terminal Gly to Ala or Ser of the GFP model substrate, which was created after removing the N-terminal SUMO domain by Ulp1 digestion, and tested the backbone cyclization of the GFPs by *Oa*AEP1 (Fig. 6) (Mikula et al. 2017). The circular form of GFP migrates faster than the linear form in the SDS-PAGE after successful backbone cyclization and can thus be used to monitor the transpeptidase activity (Iwai et al. 2001). Both Ala and Ser at the N-terminus (the P1" position) in the GFP were cyclized as efficiently as GFP with the N-terminal Gly when the P1' position is Ala (Fig. 6B). Previously, Gln and Lys was successfully used as the N-terminal residue at the P1" position for backbone cyclization of peptides by *Oa*AEP1 (Harris et al. 2015). Additionally, we tested a few hydrophobic amino-acid types at the P2" position following the N-terminal residue for backbone cyclization of the model GFP (Fig. 6B). These data suggest that the sequence requirement by *Oa*AEP1 is not strictly limited to "NGL" but could be used for other sequences, thereby widening the application of AML using *Oa*AEP1. However, the full scope of permissible amino acids has not been fully elucidated because the number of amino-acid combinations is 3.2×10^6 for the five positions in the recognition sequence.

Discussion

The ability to efficiently ligate peptide fragments and thereby introduce segmental labeling in proteins at any desired region without extensive optimizations and sequence alterations is a challenge of great importance in many areas, including protein NMR as well as protein engineering. Transpeptidases with high turnover rates can be an attractive approach because enzyme-mediated ligation would then require only a small amount of the enzyme for ligation.

In this study, we demonstrated that bacterially produced asparaginyl endopeptidase from *Oldenlandia affinis* (*Oa*AEP1) could be used to efficiently produce a segmentally isotope-labeled dArmRP, of which split fragments have a high intrinsic affinity and can be reconstituted *in vitro* (Watson et al. 2014). The high turnover and shorter recognition sequence of *Oa*AEP1 make AML an attractive alternative to SML. The minimal length required for the ligation site using *Oa*AEP1 can be as short as three residues in total, which is shorter than the required sequence of "LPXTGG" for SrtA-mediated ligation. This feature is advantageous when ligated products need to have native sequences. However, for the higher ligation efficiency, the ligation reaction must be preferred by the enzyme to the reverse reaction, such as, e.g. by the proximity effect.

The ligation site we used in this model system of dArmRP contained a sequence of "KNGL". However, the N-terminal residue of the C-terminal fragment can also be Ala or Ser instead of Gly for a *trans*-peptide reaction, thereby expanding the possible sequence combinations by *Oa*AEP1. Our experiments with a few variants at the P2" position suggest that "Leu" at the P2" position could also be replaced at least by a few hydrophobic residues such as "Val", "Ile", "Met", and "Phe" together with Ala at the P1' position (Fig. 6B). Further characterization of the specificity of *Oa*AEP1 in detail will widen the application of *Oa*AEP1 for segmental isotopic labeling as well as other protein engineering applications. AML does not require any fusions with large protein fragments like split inteins but currently requires only three residues at the C-terminus as the AEP-tag. Compared with AML, the PTS approach was successfully used to produce a native sequence after splicing, because of the high diversity of intein sequences and high specificity of the splicing reaction (Ciragan et al. 2016). Although the promiscuity of *Oa*AEP1 could be beneficial for finding an appropriate ligation site

in the native sequence, the tradeoff can be non-specific cleavages, and this might be necessary to be suppressed for, if not all, some cases.

There are now several approaches for segmental isotopic labeling. Each method still has specific requirements and restrictions. It is unlikely to have only one robust method that could be used for every situation of all proteins. Further characterization and engineering of inteins, AEPs, and SrtA could establish a general toolbox for segmental isotopic labeling of proteins and thereby segmental isotopic labeling can be used more widely for NMR studies of larger proteins, proteins with a repetitive sequence, and multi-domain proteins. Particularly transient interactions, present within large multi-domain proteins, might be difficult to observe without segmental isotopic labeling even though they may play critical roles in biological functions (Minato et al. 2017; Shiraishi et al. 2018).

Acknowledgments

This work is supported by grants from the Academy of Finland (131413, 137995). The NMR facility at the Institute of Biotechnology, University of Helsinki is supported by Biocenter Finland and Helsinki Institute of Life Science (HiLIFE). The authors thank K. Elsner, A. Hietikko, S. Jäskeläinen, J. Tommila, and E. Maschke for their technical assistance.

References

Alfarano P, Alfarano P, Varadamsetty G, Ewald C, Parmeggiani F, Pellarin R, Zerbe O, Plückthun A, Caflisch A (2012) Optimization of designed armadillo repeat proteins by molecular dynamics simulations and NMR spectroscopy. Protein Sci 21:1298-314

Antos JM, Truttmann MC, Ploegh HL (2016) Recent advances in sortase-catalyzed ligation methodology. Curr Opin Struct Biol 38:111–118

Aranko AS, Oeemig JS, Iwaï H (2013) Structural basis for protein *trans*-splicing by a bacterial intein-like domain: protein ligation without nucleophilic side-chains. FEBS J 280:3256-69

Aranko AS, Oeemig JS, Zhou D, Kajander T, Wlodawer A, Iwaï H (2014) Structure-based engineering and comparison of novel split inteins for protein ligation. Mol Biosyst 10:1023–1034

Busche AEL, Aranko AS, Talebzadeh-Farooji M, Bernhard F, Dötsch V, Iwaï H (2009) Segmental isotopic labeling of a central domain in a multidomain protein by protein *trans*-splicing using only one robust DnaE intein. Angew Chem Int Ed Engl 48:6128–6131

Ciragan A, Aranko AS, Tascon I, Iwaï H (2016) Salt-inducible protein splicing in *cis* and *trans* by inteins from extremely halophilic archaea as a novel protein-engineering tool. J Mol Biol 428:4573-4588

Dawson PE, Muir TW, Clark-Lewis I, Kent SB (1994) Synthesis of proteins by native chemical ligation. Science 266:776–779

Dorr BM, Ham HO, An C, Chaikof EL, Liu DR (2014) Reprogramming the specificity of sortase enzymes. Proc Natl Acad Sci USA 111:13343–13348

Evans TC, Benner J, Xu MQ (1998) Semisynthesis of cytotoxic proteins using a modified protein splicing element. Protein Sci 7:2256–2264

Freiburger L, Sonntag M, Hennig J, Li J, Zou P, Sattler M (2015) Efficient segmental isotope labeling of multi-domain proteins using Sortase A. J Biomol NMR 63:1–8

Gallagher C, Burlina F, Offer J, Ramos A (2017) A method for the unbiased and efficient segmental labelling of RNA-binding proteins for structure and biophysics. Sci Rep. 7(1):14083.

Guerrero F, Ciragan A, Iwaï H (2015) Tandem SUMO fusion vectors for improving soluble protein expression and purification. Protein Expr Purif 116:42–49

Harris KS, Durek T, Kaas Q, Poth AG, Gilding EK, Conlan BF, Saska I, Daly NL, van der Weerden NL, Craik DJ, Anderson MA (2015) Efficient backbone cyclization of linear peptides by a recombinant asparaginyl endopeptidase. Nat Comm 6:10199

Iwai H, Lingel A, Plückthun A (2001) Cyclic green fluorescent protein produced in vivo using an artificially split PI-PfuI intein from *Pyrococcus furiosus*. J Biol Chem 276:16548-16554

Johnson ECB, Kent SBH (2006) Insights into the mechanism and catalysis of the native chemical ligation reaction. J Am Chem Soc 128:6640–6646

Kobashigawa Y, Kumeta H, Ogura K, Inagaki F (2009) Attachment of an NMR-invisible solubility enhancement tag using a sortase-mediated protein ligation method. J Biomol NMR 3:145-50

Mao H, Hart SA, Schink A, Pollok BA (2004) Sortase-mediated protein ligation: a new method for protein engineering. J Am Chem Soc 126:2670–2671

Michel E, Skrisovska L, Wüthrich K, Allain FH (2013) Amino acid-selective segmental isotope labeling of multidomain proteins for structural biology. Chembiochem14:457-66

Mikula KM, Tascón I, Tommila JJ, Iwaï. H (2017) Segmental isotopic labeling of a single-domain globular protein without any refolding step by an asparaginyl endopeptidase. FEBS Lett 591:1285–1294

Minato Y, Ueda T, Shimada I, Iwaï H (2012) Segmental isotopic labeling of a 140 kD dimeric multi-domain protein CheA from *Escherichia coli* by expressed protein ligation and protein *trans*-splicing. J Biomol NMR 53:191-207

Minato Y, Ueda T, Machiyama A, Iwaï H, Shimada I (2017) Dynamic domain arrangement of CheA-CheY complex regulates bacterial thermotaxis, as revealed by NMR. Sci Rep 7:16462

Muir TW, Sondhi D, Cole PA (1998) Expressed protein ligation: A general method for protein engineering. Proc Natl Acad Sci USA 95:6705–6710

Muona M, Aranko AS, Raulinaitis V, Iwaï H (2010) Segmental isotopic labeling of multi-domain and fusion proteins by protein *trans*-splicing *in vivo* and *in vitro*. Nat Prot 5:574–587

Nguyen GKT, Wang S, Qiu Y, Hemu X, Lian Y, Tam JP (2014) Butelase 1 is an Asx-specific ligase enabling peptide macrocyclization and synthesis. Nat Chem Biol 10:732–738

Oeemig JS, Aranko AS, Djupsjöbacka J, Heinämäki K, Iwaï H (2009) Solution structure of DnaE intein from *Nostoc punctiforme*: structural basis for the design of a new split intein suitable for site-specific chemical modification. FEBS Lett 583:1451-1456

Otomo T, Teruya K, Uegaki K, Yamazaki T, Kyogoku Y (1999a) Improved segmental isotope labeling of proteins and application to a larger protein. J Biomol NMR 14:105−114

Otomo T, Ito N, Kyogoku Y, Yamazaki T (1999b) NMR observation of selected segments in a larger protein: central-segment isotope labeling through intein-mediated ligation. Biochemistry 38:16040-16044

Parmeggiani F, Pellarin R, Larsen AP, Varadamsetty G, Stumpp MT, Zerbe O, Caflisch A, Plückthun A (2008) Designed armadillo repeat proteins as general peptide-binding scaffolds: consensus design and computational optimization of the hydrophobic core. J Mol Biol 376:1282-1304

Reichen C, Hansen S, Plückthun A (2014). Modular peptide binding: From a comparison of natural binders to designed armadillo repeat proteins. J Struct Biol 185: 147-162

Schneider CA Rasband WS, Eliceiri KW (2012) NIH Image to ImageJ: 25 years of image analysis. Nat Methods 7:671–675

Shiraishi Y, Natsume M, Kofuku Y, Ueda T, Nakata K, Mizukoshi T, Iwaï H, Shimada I (2018) Phosphorylation-induced conformation of ß2-adrenoceptor related to arrestin recruitment revealed by NMR. Nat Commun 9: 194

Skrisovska L, Allain FHT (2008) Improved segmental isotope labeling methods for the NMR study of multidomain or large proteins: application to the RRMs of Npl3p and hnRNP L. J Mol Biol 375:151-164

Skrisovska L, Schubert M, Allain FHT (2010) Recent advances in segmental isotope labeling of proteins: NMR applications to large proteins and glycopeptides. J Biomol NMR 46:51–65

Volkmann G, Iwaï H (2010) Protein *trans*-splicing and its use in structural biology: opportunities and limitations. Mol Biosyst 6:2110-2121

Watson RP, Christen MT, Ewald C, Bumbak F, Reichen C, Mihajlovic M, Schmidt E, Güntert P, Caflisch A, Plückthun A, Zerbe O (2014) Spontaneous self-assembly of engineered armadillo repeat protein fragments into a folded structure. Structure 22:985-995

Xu R, Ayers B, Cowburn D, Muir TW (1999) Chemical ligation of folded recombinant proteins: segmental isotopic labeling of domains for NMR studies. Proc Natl Acad Sci USA 96:388-393

Yamazaki T, Otomo T, Oda N, Kyogoku Y, Uegaki K, Ito N, Ishino Y, Nakamura H (1998) Segmental isotope labeling for protein NMR using peptide splicing. J Am Chem Soc 120:5591–5592

Züger S, Iwaï H (2005) Intein-based biosynthetic incorporation of unlabeled protein tags into isotopically labeled proteins for NMR studies. Nat Biotechnol 23:736-737

Figure legends

Figure 1. Different approaches for segmental isotopic labeling (**A**) Expressed protein ligation (EPL)/Intein-mediated protein ligation (IPL) using native chemical ligation (NCL). (**B**) Protein *trans*-splicing (PTS) by split inteins. (**C**) Salt-dependent conditional protein *trans*-splicing by split inteins (CTS). (**D**) Sortase-mediated ligation (SML). (**E**) AEP-mediated ligation (AML).

Figure 2. Reaction types catalyzed by AEP (**A**) Enzymatic bi-molecular *trans*-ligation. (**B**) Backbone cyclization. (**C**) The proximity effect in enzymatic ligation.

Figure 3. AEP-mediated protein ligation in *trans*. (**A**) Schematic illustration of *trans*-ligation of two GFP variants catalyzed by *Oa*AEP1. (**B**) SDS-PAGE analysis of the time course of the ligation reaction of two GFP substrates. M stands for the molecular marker. Arrows indicate the bands of substrates and the ligated product. (**C**) The estimated ligation yield *versus* time.

Figure 4. Ligation of the designed armadillo repeat fragments by *Oa*AEP1
(**A**) Schematic representation of the production of segmentally labeled dArmRP by protein-ligation catalyzed by *Oa*AEP1. (**B**) SDS-PAGE analysis of *trans*-ligation between the N-terminal YM and C-terminal M_2A from dArmRP. 0h and 3h indicate hours after the addition of *Oa*AEP1, O/N and M stand for overnight incubation and molecular marker, respectively. Arrows indicate C- and N- terminal fragments and full-length dArmRP as the ligated product. (**C**) SDS-PAGE analysis of the segmentally labeled dArmRP after anion-exchange chromatography.

Figure 5. Segmental labeling of dArmRP (**A**) A cartoon model of the structure of the segmentally labeled dArmRP (YM_3A) in which [^{13}C, ^{15}N]-labeled YM region and M_2A unlabeled regions are colored in red and dark blue, respectively. (**B**) Sequences of N- and C-terminal fragments in red and black, respectively; [^{13}C, ^{15}N]-labeled N-terminal fragment for YM domains containing a C-terminal “KNCL” recognition tag for AEP

(underlined and in italics); C-terminal fragment for M_2A domains with the N-terminal "GL" recognition tag for AEP (underlined); the expected sequence of the ligated product. (**C**) [^{1}H, ^{15}N]-HSQC spectra of uniformly labeled (left) and segmentally [^{13}C, ^{15}N,]-labeled dArmRP produced by AML (right).

Figure 6. The promiscuity of *Oa*AEP1 at the N-terminal recognition sequence

(**A**) Schematic representation of the substrate (GFP) and reaction catalyzed by *Oa*AEP1. The C-terminal P1-P3 and P1'-P3' sites of the C-terminal propeptide are indicated. P1" and P2" denote the N-terminal residues that replace the P1' and P2' residues after cleavage of the C-terminal propeptide. The amino acid sequence corresponding to the natural substrate of kalataB1 are shown on the top. The amino acid types reported in this work and previous reports are shown below ([a]Harris et al. 2015 and [b]Mikula et al. 2017). (**B**) SDS-PAGE analysis of backbone cyclization of GFP substrates by *Oa*AEP1. Variants of the N-terminal recognition sequence at the P1" or P2" site are indicated above the gels; 0h, 0.5h, 1h, and 2h indicate the number of hours after addition of the enzyme; M stands for molecular weight marker, and arrows indicate the bands of linear and circular GFPs.

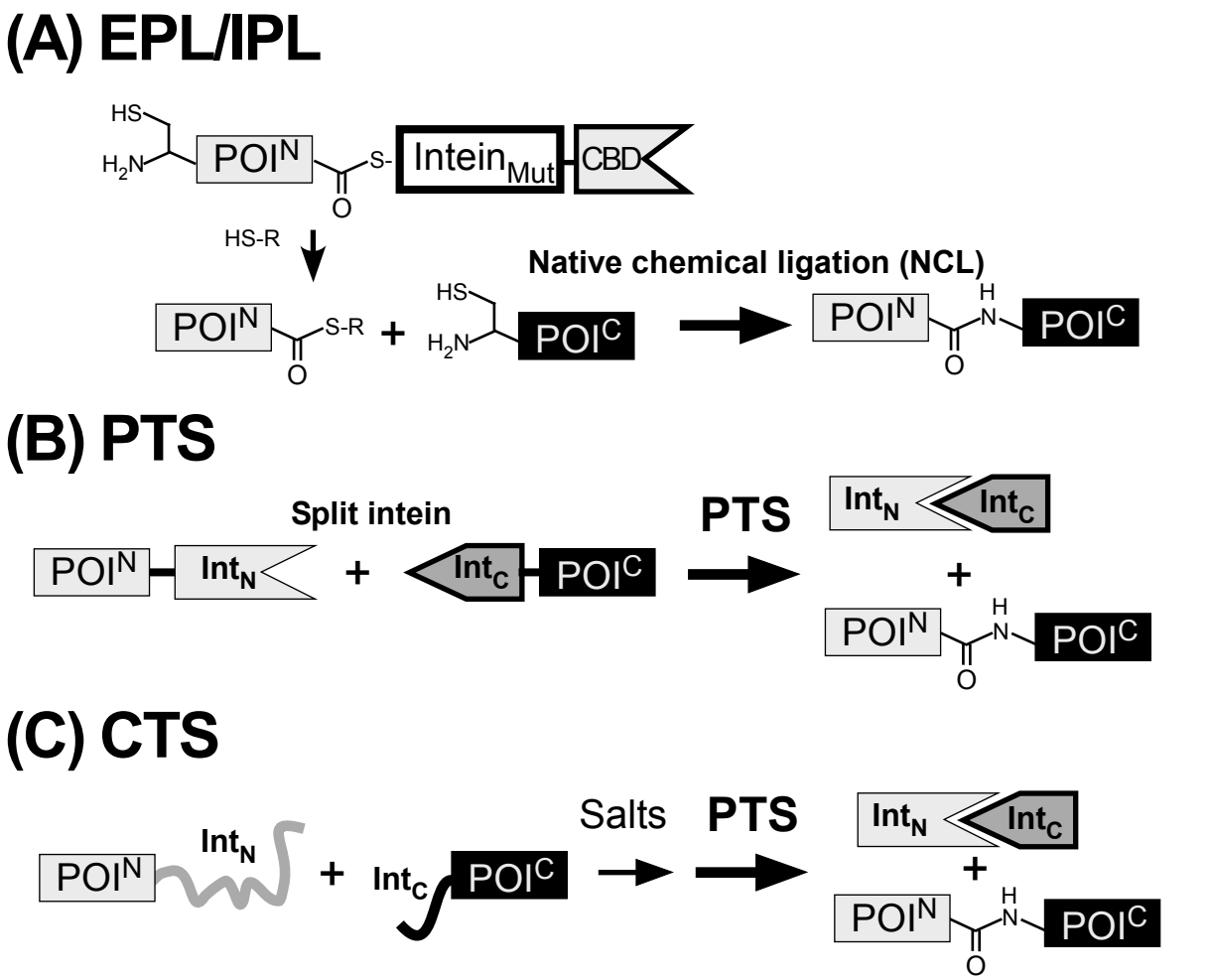


Fig. 1

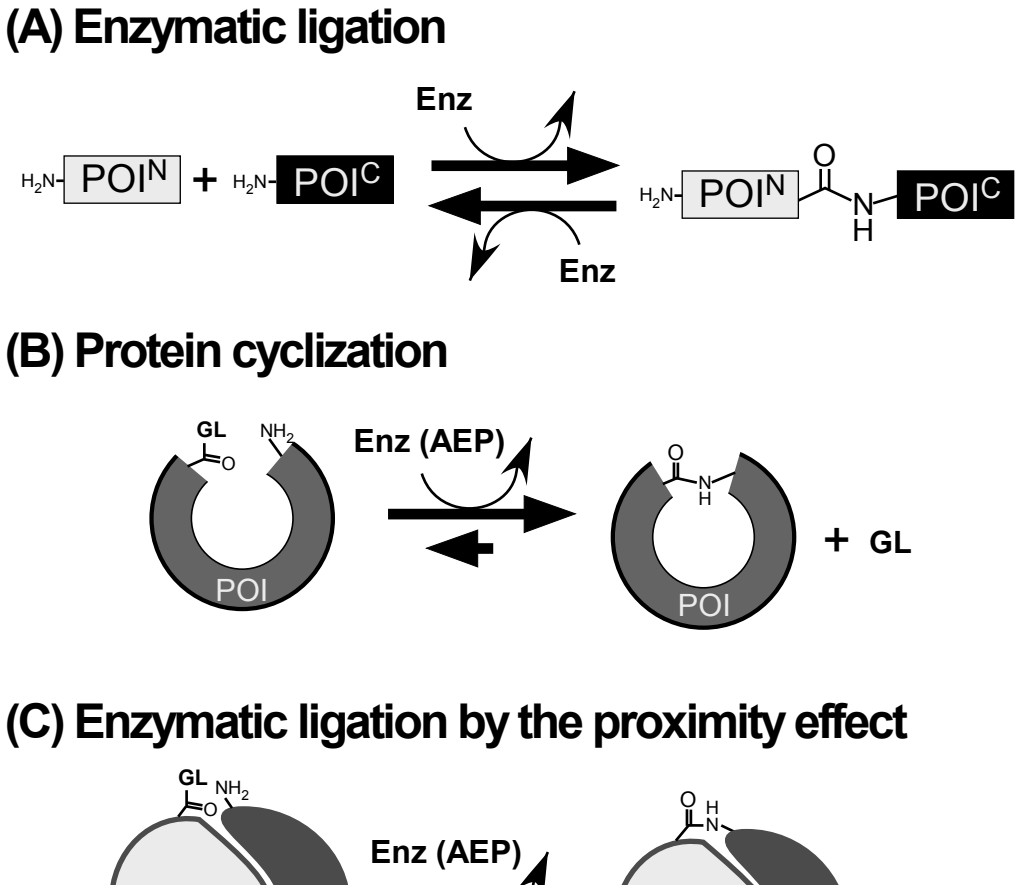


Fig. 2

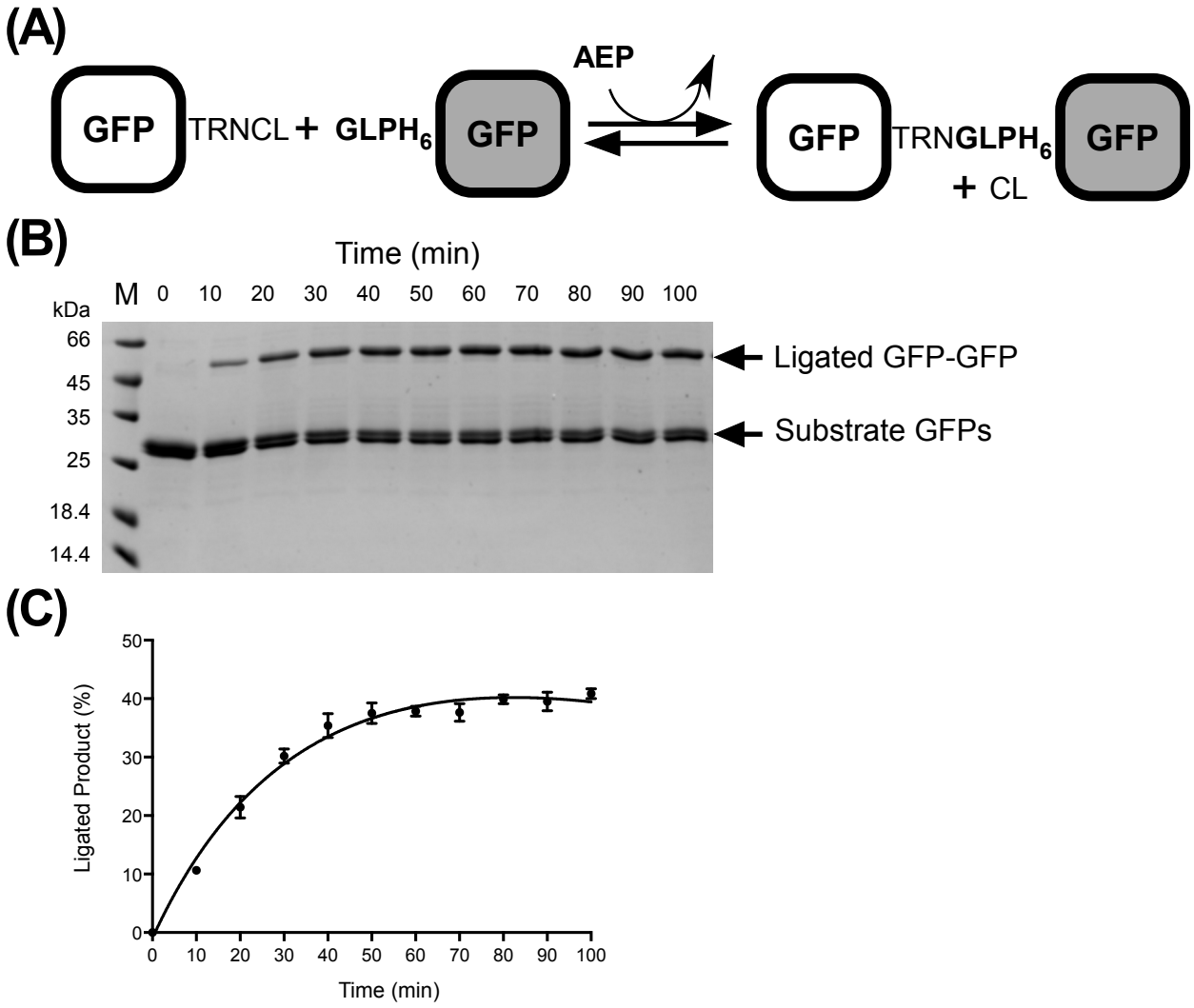


Fig. 3

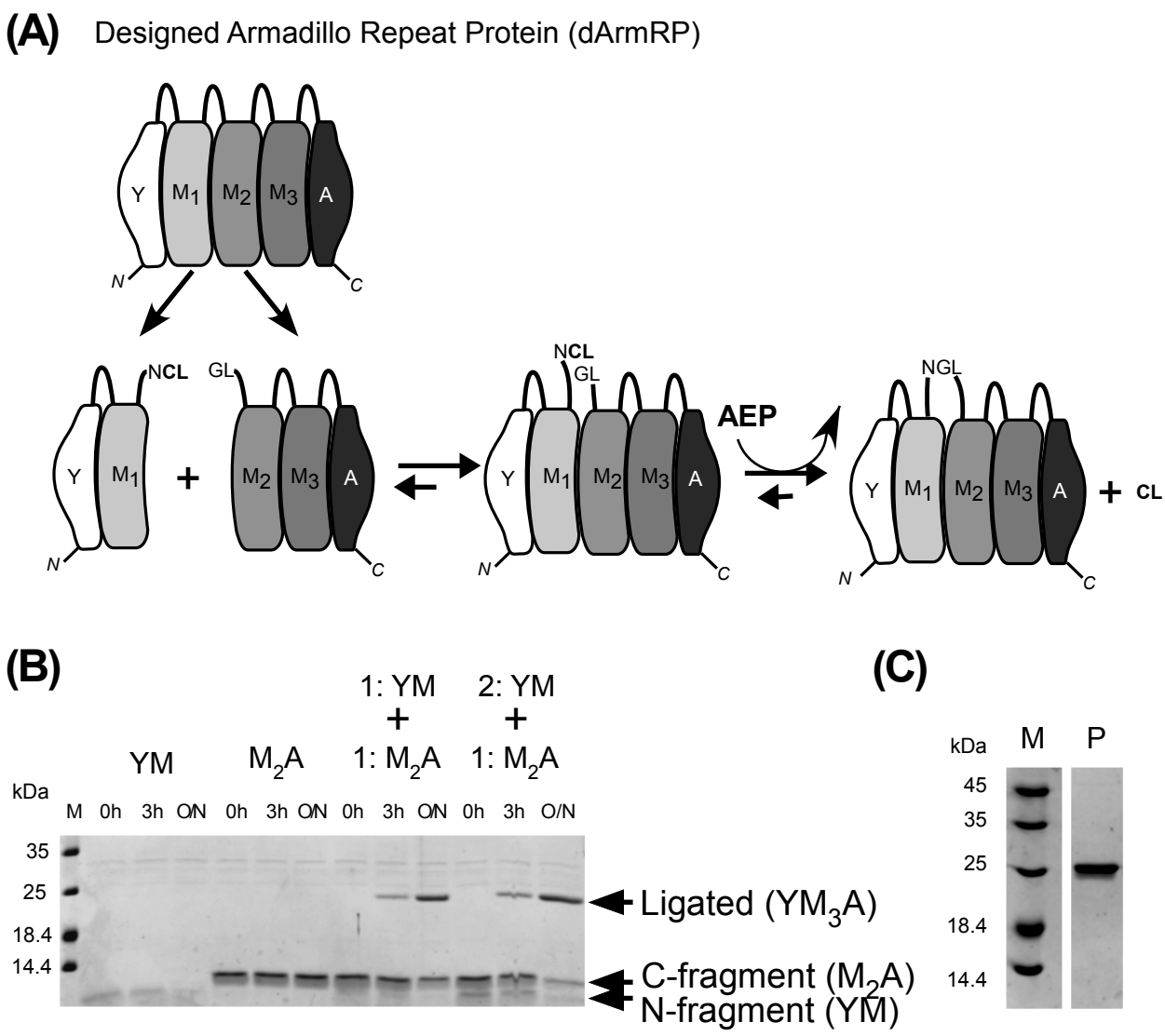
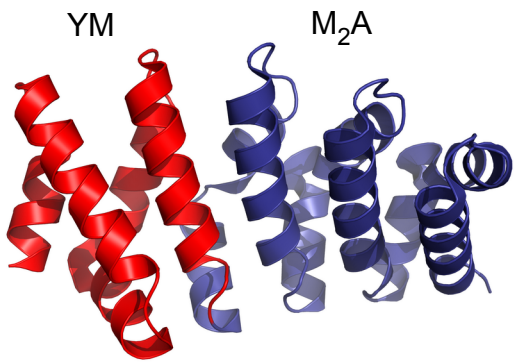


Fig. 4

(A)



(B)

N-terminal fragment

```
Y: SELPQMTQQLNSDDMQEQLSATRKFSQILSD
M: GNEQIQAVIDAGALPALVQLLSSPNEQILQEALWALSNIASGKNCL
```

C-terminal fragment

```
M: GLNEQIQAVIDAGALPALVQLLSSPNEQILQEALWALSNIASG
M: G-NEQIQAVIDAGALPALVQLLSSPNEQILQEALWALSNIASG
A: G-NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKLQSH
     EFHHHHHH
```

Ligated product

```
Y: SELPQMTQQLNSDDMQEQLSATRKFSQILSD
M: G-NEQIQAVIDAGALPALVQLLSSPNEQILQEALWALSNIASGKN
M: GLNEQIQAVIDAGALPALVQLLSSPNEQILQEALWALSNIASG
M: G-NEQIQAVIDAGALPALVQLLSSPNEQILQEALWALSNIASG
A: G-NEQKQAVKEAGALEKLEQLQSHENEKIQKEAQEALEKLQSH
   EFHHHHHH
```

(C)

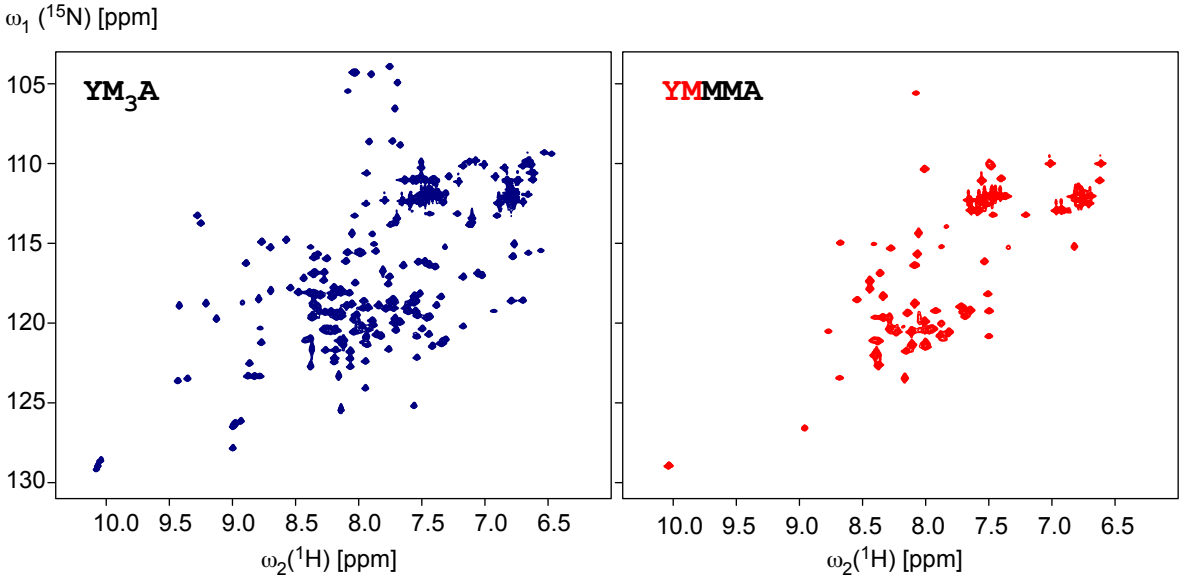


Fig. 5

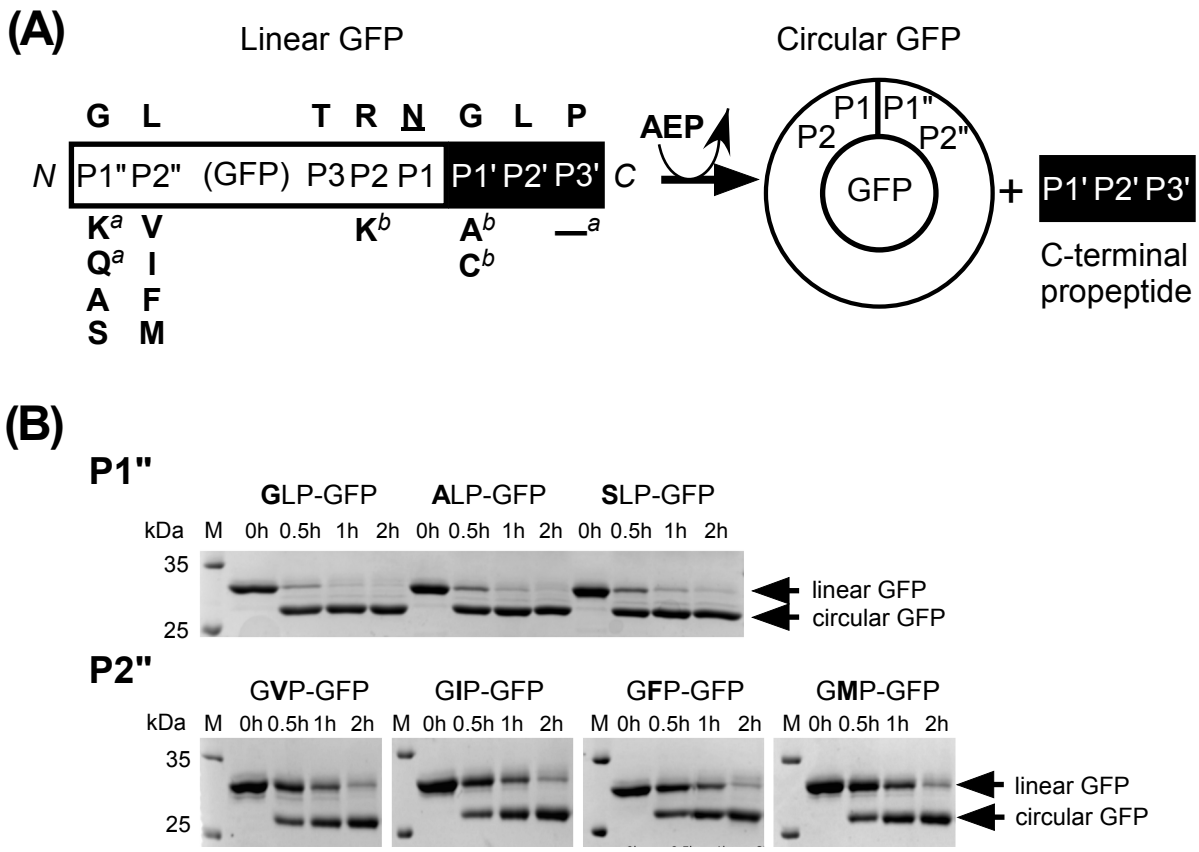
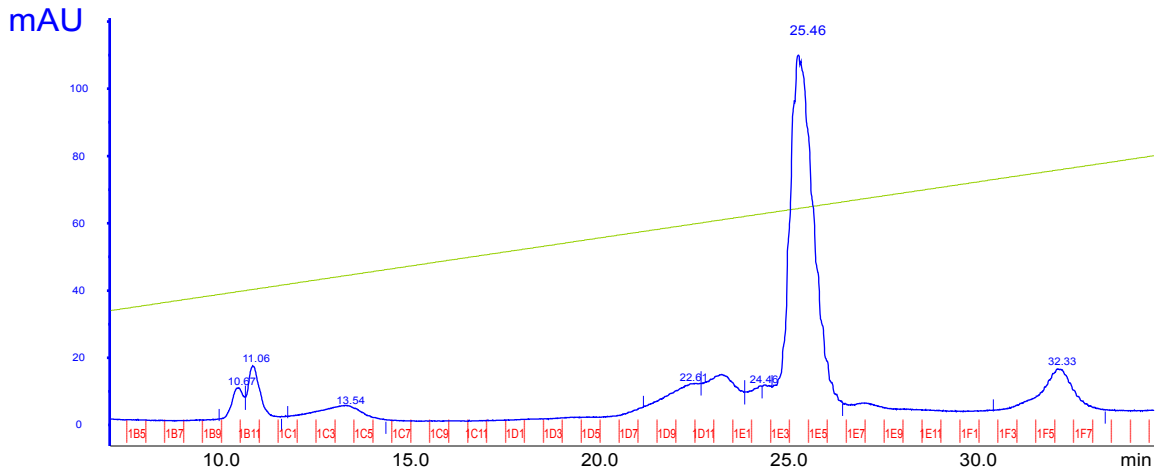
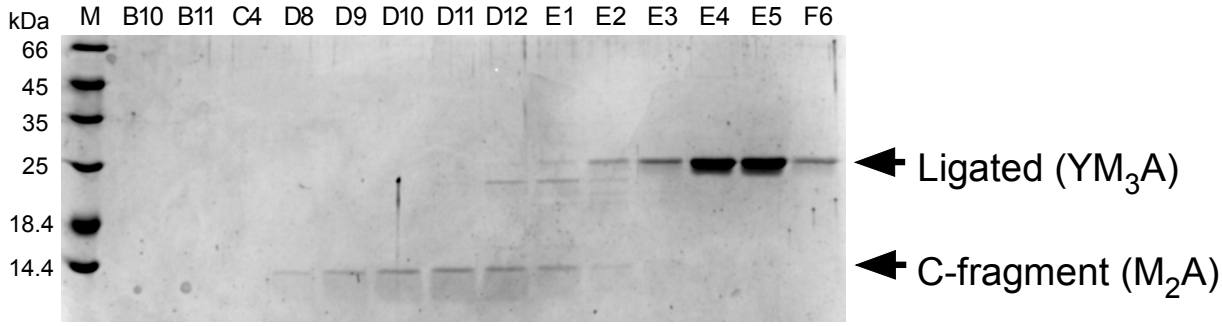


Fig. 6

(A)



(B)



Supplementary Figure 1: Purification of the ligated product from the reaction mixture by anion exchange chromatography. (**A**) Chromatogram from anion exchange chromatography using MonoQ 5/50 GL column (GE Healthcare). (**B**) SDS-PAG analysis of the elution fractions by a linear gradient of 250-800 mM NaCl in 20 mM Tris, pH 8.0.